An evaluation framework for groups' clustering algorithms in social networks.

The use case of a Meetup dataset of older adults.

Marcelo RODAS BRITEZ [a,1] , Davide LISSONI [a] and Maurizio MARCHESE [a,2]
[a] *University of Trento, Trento, Italy*

Abstract. Older adults are increasingly using Social Networks Sites to support their social interactions. Moreover, the popularity of such network sites, the availability of datasets and recent progress in the computing systems and machine learning areas have made social network analysis a current important area of research. In this context, our current research aim is to investigate how these technological platforms are affecting the lifestyle of older adults. In this paper, we propose a Java oriented framework that can assist researchers in this area in the analysis of social networks groups, specifically in the comparison of user's groups creation from existing cluster-based algorithms. To validate the proposed framework we use a dataset extracted from the Meetup social network - a website providing members software services to schedule events using a common platform. For our study, we filtered the data for the specific group of older adults. The framework proposes several ways of evaluating the quality of the data, and is extensible to other clustering algorithms and evaluation metrics. Currently, we have tested our framework with the following well known clustering algorithms: k-means, fuzzy k-means, and affinity propagation. We report some preliminary results obtained by using the proposed framework and the above clustering algorithms using the extracted Meetup dataset.

Keywords. Clustering algorithms, Recommendation Systems, Online Social Networks, Technologies for older adults

Introduction

Group recommender systems are based on the idea that people usually prefer to do activities in groups: they like to share news, discuss in groups, and perform activities in a social environment. For example, it is much more pleasant to exchange opinions among friends on news that everyone knows, than talking about an event that only a single person has seen. In fact, these types of recommender systems are interesting because events and activities are used at least as often by groups as by individuals [6].

Clustering entities by features is a well known problem in computer science [18]. Many algorithms have been proposed to perform this operation. However, typically the use of different clustering algorithms leads to different results i.e. different clusters/groups. Each clustering algorithms has some predefined parameters strictly based

[1] marcelo.rodasbritez@unitn.it
[2] maurizio.marchese@unitn.it

on the implementation of the algorithms. For example, the distance mean of the cluster instances from the cluster centroid is used in k-means evaluation, or the similarity measure between the cluster instances and its exemplars is used in the Affinity Propagation algorithm.

In this paper, we present a framework designed to provide pattern analysis of groups' clustering algorithms in order to compare their results in a systematic way, facilitating the developers in the selection of the best clustering algorithms that fit better their specific requirements and use cases. The proposed framework is based on Java patterns and it is extensible: new algorithms, statistics and quality measures can be easily added. In the present work, three different cluster algorithms were implemented and initially analyzed: K-means [19], Fuzzy K-means [4], and Affinity propagation [5].

Our framework analysis focuses on groups' clustering in online social networks. First, we perform older adult group clustering based on affinity to create social groups. The evaluation of the groups generated by clustering algorithms is based on the comparison of the created groups with existing groups in a Meetup dataset. A priori, evaluation about new groups created by different clustering algorithm can then lead social researchers to analyze the relations and distribution of data generated by the social interactions.

The paper is structured as follows: In Section 1, we present the context and motivation for our work. Section 2 briefly presents current approaches and methodologies in the domain of evaluation of clustering algorithms. In Section 3, we present the proposed evaluation frameworks and its main design principles. In Section 4, we describe the dataset used in our research to provide some examples of the typical evaluations and analysis that can be done with our evaluation framework. Conclusions and future work close the paper.

1. Groups' clustering for older adults

The underlying motivation for building this framework is to understand how recommender systems could assist older adults with the activities they do. For instance, friends' group recommendation is a recommender system approach that takes into consideration social information by creating and recommending to an older adult user of the system the participation/enrollment to a number of personalized groups. We focus on using these algorithms in social network sites (SNS) since they are becoming always more popular also in the demographic sector of older adults. The social interactions that happen in such SNS could be assessed by recommendation systems that provide useful and contextual information to the users, keeping their attention and interest, while satisfying the conditions (preferences and needs) and the constraints of the users (physical or cognitive limitations), which are specific to this demographic sector.

Researchers studied the declining wellbeing of older adults, which is related to problems like isolation, depression, inactivity, functional decline, and fear of leaving home [16]. In addition, this population sector is growing and the current health system has difficulties addressing appropriate solutions to their needs [17].

In the context of our current research - aimed to investigate and try to understand how technological platforms can affect and maybe support the active lifestyle of older adults - we are interested in studying the relationships and behaviors that emerge in

SNS. The long-term objective of our study is to support a sustainable active life-style for older adults. To do so, we propose analyzing how groups are generated, in our particular context, and to study how the clustering algorithms represent the existing relationships of groups. This will provide the basis for recommending appropriate social and physical activities. We expect that this initial stage allows us to understand better the preferences of the users and eventually use this knowledge to provide better groups' and activities' recommender systems for older adults.

2. Related work

For the development of our proposed framework, we first studied the current techniques of evaluation of clustering algorithms and how existing software libraries support such evaluation.

2.1. Evaluation of clustering algorithms

A widely used technique in the evaluation of clustering algorithms is to compare the results of the algorithms with an external criterion. This technique consists in obtaining a real dataset of users and groups as the external criterion: usually, this dataset is called the "ground truth" dataset and the related evaluation technique the "ground truth" approach. The idea then is to test various algorithms and identify the one(s) that produce the most similar results with respect to the "ground truth" dataset. This technique introduces the problem of defining a similarity measurement, and actually, there are various ways to define cluster similarity. One can use similarity measures based on the distances between cluster centroid [11]. Also, one could instead calculate clusters similarity by using the Jaccard similarity measure between cluster instances pairs [13]. Another classic way to calculate clusters similarity is to count how many instances two clusters have in common.

Another measurement technique is to use the external criterion without using a similarity measure by calculating the purity measure[14]. The purity measure is evaluated on how well clusters matches with a predefined set of classes.

Another standard way to measure the overall quality of a clustering algorithm performs is to measure the performance in terms of the Precision, Recall, and F-measure as described in [14], by identifying the following variables of the classification: true positive (*TruePositive*) as the correctly assigned clusters, false positive (*FalsePositive*) as the incorrectly assigned clusters, and False Negative (*FalseNegative*) as the not assigned clusters when it should be assigned. The formulas are described as the following:

- $Precision = \frac{TruePositive}{TruePositive+FalsePositive}$
- $Recall = \frac{TruePositive}{TruePositive+FalseNegative}$
- *F-measure*$_\beta = \frac{(1+\beta^2)(Precision*Recall)}{\beta^2 Precision+Recall}$, where β weighs the importance of Precision and Recall (β higher than 1 weighs more the Recall).

These measures are used to evaluate classification supervised learning algorithms, and they are built over the confusion matrix, which is a matrix that shows how many users were incorrectly assigned [14]. Since some clustering algorithms use also unsupervised learning approaches, in these cases to generate the confusion matrix there is the need to use ground truth groups as the corrected prediction. Ground truth is then used to

define the "predefined classes"(groups), while the clusters generated represent the "actual classes". The confusion matrix is created through the comparison between the predefined cluster instances and the actual clusters. Precision is calculated as the fraction of pairs correctly put in the same cluster, Recall is the fraction of actual pairs that were identified, and F-measure is the harmonic mean of the precision and the recall [14].

Another way of comparing two clusters, based on information theory, is called "variation of information" (VI) [7]. This measurement defines a distance between two partitions of the same data set, by measuring the amount of information lost and gained in changing from one cluster to another one.

2.2. Existing libraries

A number of free and open source machine learning Java projects, focused on data mining algorithms, exists and are used to help developers during the implementations of such algorithms. Apache Mahout [1], for example, is a free Apache Software Foundation platform that makes available different scalable machine learning algorithms focused primarily in the areas of collaborative filtering built on the top of Apache Hadoop. Another popular Java framework is WEKA: WEKA is a collection of machine learning algorithms for data mining tasks [2]. Regarding algorithms evaluations, both WEKA and Apache Mahout contains implementations of some evaluation measures.

Namely, Mahout provides only limited cluster quality evaluation, while WEKA has three different ways to measure the quality of a cluster:

1. The percentage of instances contained in each cluster;
2. The possibility to evaluate clusterings on separate test data if the cluster representation is probabilistic (e.g. for Expectation Maximization);
3. Classes to clusters evaluation: in this mode WEKA first ignores the class attribute and generates the clustering. Then during the test phase, it assigns classes to the clusters, based on the majority value of the class attribute within each cluster. Then it computes the classification error, based on this assignment and also shows the corresponding confusion matrix.

In comparison with the many features that the existing frameworks can offer, the main goal of our framework is to help the developers in the evaluation and comparison of clusters using an extensible set of personalized quality measures. Specifically, our framework aims to enrich the already existing functionalities by providing an extensible platform to easily add different cluster evaluation measures as well as comparison methodologies.

3. A Framework for analysis of social network sites groups' clustering algorithms

A framework is designed as a reusable and extensible architecture for various application domains [10]. Usually, developers want higher productivity and shorter time-to-market for the development of object-oriented applications, and these goals are achieved through a good design and reusable architectures.

Our proposed framework is built on top of the Apache Mahout library version 0.12.2, using Java 8 programming language. We use design patterns to structure the implemen-

tation of the three initially selected clustering algorithms: K-means, Fuzzy K-means and Affinity Propagation. The selected clustering algorithms are only a first set of algorithms we have started to use in our research. The proposed framework will allow to add novel and more recent clustering algorithms when needed.

To achieve a good design and reusable architecture, in our overall design, we choose to adhere to the following principles:

- Extensibility: the framework should be extensible. This means that a user can add functionalities to the frameworks without changing the existing code.
- Inversion of control: the framework maintains the control of the application life-cycle.
- Interface and class segregation: the framework should separate different functionalities into different interfaces and different entities in different classes.
- Dependency inversion: high-level framework components shall perform their functions using lower-level framework components, through the interfaces exposed by the latter.

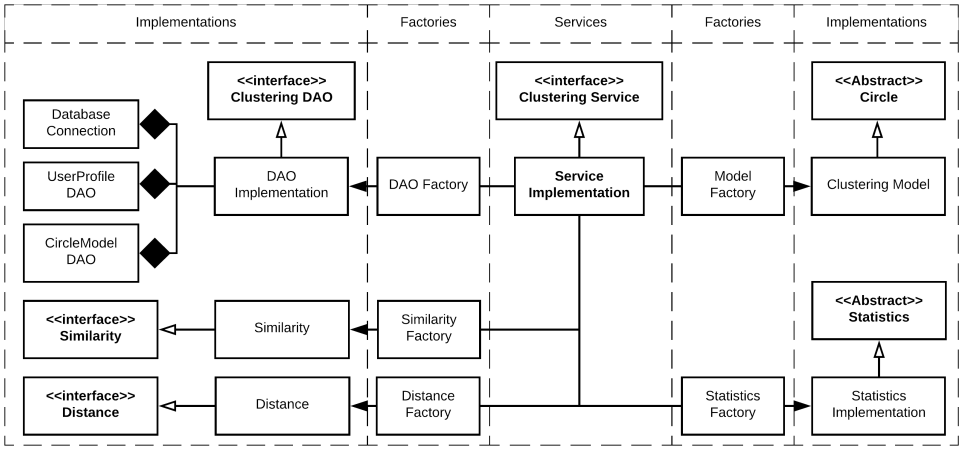

Figure 1. Framework architecture.

The architecture in Figure 1 shows the different layers of implementation chosen in the design of the framework. The use of data access objects (DAO), factories, abstract classes, and interfaces, are the main design patterns used to achieve the principles mentioned above.

Regarding the database choice, relational databases were not designed to cope with the scale and agility challenges that face modern application [8]. A dataset graph structure, on the other hand, allows to perform efficient, constant-time operation, and allow to traverse a big amount of connections per second per core [12].

Since, for our research purpose, we had to handle social information, our database choice has been oriented on graph structure databases and in particular to OrientDB database version 2.2.18. OrientDB is the first Multi-Model Open Source NoSQL DBMS that combines the power of graphs with documents, key/value, reactive, object-oriented and geo-spatial models into one scalable, high-performance operational database [9]. Our choice of OrientDB database is motivated not only by its graph structure but also by

the presence of Java API. However, since the framework is extensible, other databases could be used by simply adding additional connections and data access objects classes.

4. Tests and results

We analyze the social information domain, including people and groups. We classify the evaluation of this domain in two types: evaluation with internal criterion, and evaluation with external criterion.

The implemented evaluations using internal criterion are execution time and quality of the groups. Further, the implemented evaluations with external criterion are the correctness of groups, and the pairwise comparison.

Distances between cluster members, the density of the data space, statistical distributions are just some of the possible parameters utilized in order to evaluate the quality of either the existing or created clusters. All these performance measures, are valid clusters evaluation measures but are dependent on the used clustering algorithm.

4.1. Datasets

For our study, we obtained a testing dataset from the Meetup [3] social network API [3]. Meetup is an event based social network that facilitates hosting events in various localities around the world. Users are subscribed to Meetup mainly to organize or participate in meet-ups. Furthermore, Meetup users can create groups manually or subscribe to existing ones.

We created two different datasets (including users, groups, and related tags) based on the users of the Meetup's base group called "60+ Happy Hour":

- Meetup 1 is the dataset of the users and groups from the base group and related users and groups (all related members of the related members of the base group).
- Meetup 2 is the dataset of the users and groups from the related users and groups of the base group (only the first layer of related members of the base group).

Table 1 shows some general statistics of the two created datasets. Please note that in both datasets there are more groups than users: in fact in the Meetup social network service any single user can choose to participate to more than one group.

Table 1. Meetup dataset data distribution.

General Statistics	**Meetup 1**	**Meetup 2**
Number of users	2111	489
Number of tags	4340	1248
Average tags per user	26.01	22.99
Number of groups	3767	942
Average groups size	5	5

[3] https://www.meetup.com/

4.2. Evaluation Measures: Quality, Correctness, and Pairwise Comparisons

The "quality" of a group is defined by taking into account the parameters over which the data-point have been grouped. Since our parameters for grouping are the tags, we measure the quality by calculating the percentage of tags of the group that belongs also to the users. We choose to evaluate the clusters by using "internal" and "external" criteria, namely:

- internal criterion means that the parameters used in the evaluation of the cluster quality are derived from the clusters themselves, hence obtained without introducing external factors.
- external criterion means that the parameters used in the evaluation of the cluster quality come from a ground truth. In this case, the evaluation of the cluster quality is based on a comparison between the cluster generated and the ground truth group.

Another measure we use in our evaluation framework is "clusters user correctness": it describes the percentage of users of the algorithm's generated cluster that are present in a ground truth group. The "clusters tags correctness" percentage instead describe the percentage of tags of the generated clusters that are present in a ground truth group.

Then, the "pairwise comparison" measure is obtained by calculating the ability of the algorithm to classify pairs of instances or tags correctly. A pair of instances is classified as correct when both the elements of the pair present in a ground truth group are also present in a cluster generated. We say then that the algorithm produces 100% correct results when each instances pair contained in all the ground truth groups are also contained in the clusters generated. Regarding tags instead, the comparison logic is the same but based on cluster tags pairs. In pairwise comparison, we evaluate the algorithm using precision, recall, and f-measure (β = 0.5) evaluation measures as defined in Section 2 and in [14].

The precision, in our study, shows that we are comparing the results of the algorithms and the ground truth as a classification problem of the users. So, precision means the percentage of assigned groups that correspond to the ground truth within all assigned groups. On the other hand, recall means the percentage of assigned groups that correspond to the ground truth within all groups of the ground truth.

4.3. Internal Evaluations

The first internal evaluation we have considered is the collection of the execution time of the three algorithms. Our results tell us that the execution time is influenced more by the number of users' tags in the dataset than the number of users for all analyzed clustering algorithms.

Then we have focused our analysis on the internal quality evaluation and we report our results in Figure 2 obtained using the Meetup 2 dataset. In this figure, the Y-axis group with ranges the qualities of groups, and the X-axis show the percentage of the quality groups within the ranges over the total amount of groups. We think that the overall internal evaluation quality does not go beyond 30% because the average overlap over the user's tags and the group's tags is around 30% in our dataset. Our results in this evaluation show that the Fuzzy k-means algorithm provides slightly better quality results over the whole spectrum of qualities of groups.

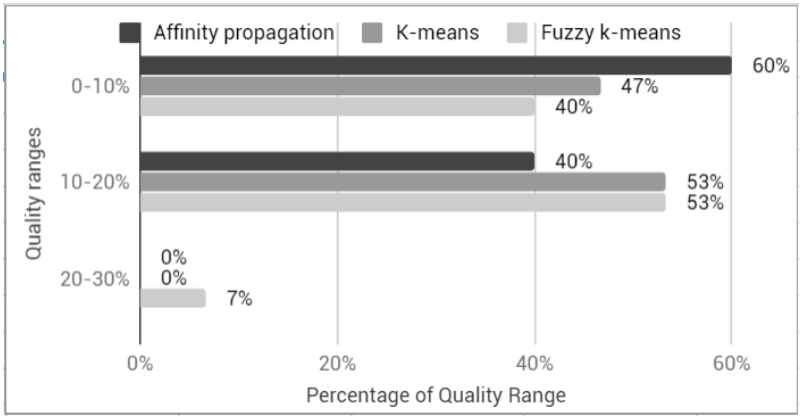

Figure 2. Cluster quality distribution resulted by running Affinity Propagation, K-means and Fuzzy K-means algorithms on the Meetup 2 dataset.

4.4. External Evaluations

The results in the Tables 2 and 3 show that the best results we have obtained from the Meetup 2 dataset using the external criterion of the clusters are provided by the Affinity Propagation algorithm.

In Table 2 we notice that the Correctness is better in terms of tags than in term of users. In our dataset, this is related to the fact that the tags of the clusters are calculated out of the tags of the users and not based on an explicit description of the ground truth. These results are correlated with better precision as shown in the table 3.

In Table 3 we notice that in terms of pairwise comparison the Recall is significantly low in all cases. This relates to the fact that all algorithms are not covering well most of the relevant results. Also, we notice that the recall using tags pairs is somehow better (but still low) than the recall using users pairs because the tags of the groups are calculated base on the users of the groups, thus improving the precision.

Table 2. Results on cluster user and tags correctness made by using the external criterion for the clusters generated by the three algorithms implemented on top of the framework using the Meetup 2 dataset.

Algorithm	**Cluster users' correctness**			**Cluster tags' correctness**		
	AVG	**MAX**	**MIN**	**AVG**	**MAX**	**MIN**
Affinity propagation	56.45%	79.17%	13.64%	95%	99.36%	89.44%
K-means	52.28%	85.71%	24%	94.20%	99.76%	88.78%
Fuzzy K-means	38.50%	100%	12.50%	60%	100%	49.92%

4.5. Datasets comparison

The proposed evaluation framework also allows a comparison between different datasets, so we could see the similarities and differences of the clustering algorithms on different type of datasets.

Table 3. Results on pairwise users and pairwise tags comparisons made by using the external criterion of the clusters generated by the three algorithms implemented on top of the framework using the Meetup 2 dataset.

Algorithm	Pairwise users comparison			Pairwise tags comparison		
	Precision	**Recall**	**F-measure**	**Precision**	**Recall**	**F-Measure**
Affinity propagation	62.04%	3.82%	7.21%	99.38%	17.22%	29.35%
K-means	62.24%	4.13%	7.76%	99.22%	15.28%	26.49%
Fuzzy K-means	19.50%	0.01%	0.19%	80.02%	0.70%	1.39%

The results in Table 4 show that the precision and recall of the classification are higher in Meetup 2. Also, the recall is significantly lower in Meetup 1. In general the low recall in all investigated algorithms means that they are missing to discover a large number of possible groups.

Analyzing Tables 1 and 4, the datasets Meetup 1 and Meetup 2, have proportionally the same data distribution. However, since the Meetup 2 dataset is composed of users coming from a common group, we expect that, by running clustering algorithms on this dataset, we would obtain higher evaluations of precision and recall. The evaluation study confirms this assumption.

Table 4. Average pairwise evaluation using the external criterion of the clusters generated by the three algorithms implemented for the two analyzed datasets.

Dataset	Pairwise users comparison			Pairwise tags comparison		
	Precision	**Recall**	**F-measure**	**Precision**	**Recall**	**F-Measure**
Meetup 1	14.22%	0.38%	0.75%	95.23%	1.99%	3.90%
Meetup 2	54.76%	5.84%	10.30%	99.01%	16.19%	27.82%

5. Discussion and Conclusions

We have developed and implemented an extensible Java framework with the aim of giving developers of clustering algorithms the opportunity to implement, evaluate and compare their algorithms. The framework is designed to execute data mining algorithms on users' data. Moreover, the framework architecture allows to extend to different algorithms, evaluation metrics and different domains.

The main contribution of our work is the definition of an evaluation environment based on the current best practices in terms of development and capable to support the evaluation and comparison of clustering algorithms in a systematic way.

To test the framework we have used two datasets obtained from a specific use case related to older adults in the Meetup social network web site. We thus used the proposed evaluation framework to compute and compare a number of quality metrics using three state of the art clustering algorithms. The preliminary results highlighted the ability of the framework to produce comparable quality measures and algorithms evaluations. Moreover, the framework structure gives us the possibility to execute, evaluate and compare the selected (and implemented) algorithms, also by quickly changing their execution parameters. More importantly, in fact, we were able to draw some conclusions about the different algorithms behavior and their results for the specific datasets.

For instance, our evaluation framework helped us to understand, that Affinity Propagation algorithm provides better results if we analyze its quality performances using an external criterion instead of an internal one. Moreover, we found that all the three implemented algorithms are giving relatively precise groups but they are missing lots of existing groups present in the ground truth dataset.

In the context of our current research - aimed to investigate and try to understand how technological platforms can affect and maybe support the active lifestyle of older adults - we are interested in studying the relationships and behaviors that emerge in Social Network Sites. We want to use the proposed evaluation framework, to evaluate how different clustering algorithms are capable to capture and represent existing relationships of groups in real social networks. This will provide the basis for recommending participation to appropriate interest groups to older adults and - in a different, but related thread of research - potentially interesting social and physical activities.

References

[1] Apache mahout. http://mahout.apache.org/. Accessed: 2018-03-13.

[2] Weka 3: Data mining software in java. https://www.cs.waikato.ac.nz/ml/weka/. Accessed: 2018-03-13.

[3] Meetup API. https://www.meetup.com/meetup_api/. Accessed: 2018-03-13.

[4] Apache Mahout documentation. Fuzzy k-means. https://mahout.apache.org/users/clustering/fuzzy-k-means.html. Accessed: 2018-03-13.

[5] Brendan J Frey and Delbert Dueck. Clustering by passing messages between data points. science, 315(5814):972-976, 2007.

[6] Anthony Jameson and Barry Smyth. Recommendation to groups. In The adaptive web, pages 596-627. Springer, 2007.

[7] Marina Meila. Comparing clusterings-an information based distance. Journal of multivariate analysis, 98(5):873-895, 2007.

[8] MongoDB. Nosql database explained. https://www.mongodb.com/nosql-explained. Accessed:2018-03-13.

[9] OrientDB. Why a multi-model database? https://orientdb.com/multi-model-database/. Accessed: 2018-03-13.

[10] Wolfgang Pree. Framework development and reuse support. Burnett et al.[5], 1995.

[11] Julien Soler, Fabien Tence, Laurent Gaubert, and Cedric Buche. Data clustering and similarity. In FLAIRS Conference, 2013.

[12] Poornapragna Malamandi Suresh. Database as a graph store, 2016.

[13] Guadalupe J Torres, Ram B Basnet, Andrew H Sung, Srinivas Mukkamala, and Bernardete M Ribeiro. A similarity measure for clustering and its applications. Int J Electr Comput Syst Eng, 164-170, 2009.

[14] Christopher D. Manning, Prabhakar Raghavan, and Hinrich SchÃijtze. 2008. Introduction to Information Retrieval. Cambridge University Press, New York, NY, USA.

[15] Xiwang Yang, Harald Steck, and Yong Liu. Circle-based recommendation in online social networks. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 1267-1275. ACM, 2012.

[16] Schutzer, K. A., & Graves, B. S. (2004). Barriers and motivations to exercise in older adults. Preventive medicine, 39(5), 1056-1061.

[17] Vroman, K. G., Arthanat, S., & Lysack, C. (2015). "Who over 65 is online?" Older adults' dispositions toward information communication technology. Computers in Human Behavior, 43, 156-166.

[18] Brucker, Peter. "On the complexity of clustering problems." Optimization and operations research. Springer, Berlin, Heidelberg, 1978. 45-54.

[19] Hartigan, John A., and Manchek A. Wong. "Algorithm AS 136: A k-means clustering algorithm." Journal of the Royal Statistical Society. Series C (Applied Statistics) 28.1 (1979): 100-108.